# MAT 303 Module One Problem Set Report

Multiple Regression

Nicholas Kreuziger

nicholas.kreuziger@snhu.edu

Southern New Hampshire University

## Introduction

The data set being explored is from the 1974 Motor Trend US Magazine. This data set touches upon fuel consumption and 10 aspects of automobile design and performance for 32 automobiles. The models of these cars are from the years 1973-1974. The results may be used to determine what variables are most important for a car’s fuel economy. Car makers are interested in the variables that contribute the most to Fuel Economy. The analyses being run will be a Correlation Analysis and a Multiple Regression Analysis.

## Data Preparation

There are 11 variables in this data set in a 32 Row and 12 column table. The important variables identified as the predictor variables are weight (wt.) and horsepower (hp). Weight is measured in units of 1,000 lbs. If the Weight variable is 2.5 it represents 2,500 lbs. Horsepower, when ft\*lbf and rpm is available, is calculated by the equation . The unit for horsepower is ft∙lbf/min, 3 horsepower of force moves 3 lb. per foot of force per minute.

## Multiple Regression Model

### Correlation

When viewing a scatter plot of Fuel Economy against Weight you can see a negative linear relationship. As Weight increases Fuel Economy decreases.

Chart, scatter chart

Description automatically generated

When viewing a scatter plot of Fuel Economy against Weight you can see a negative linear relationship. As Weight increases Fuel Economy decreases.

**Chart, scatter chart

Description automatically generated**

When viewing a scatter plot of Fuel Economy against Horsepower you can see a negative linear relationship. As Horsepower increases Fuel Economy decreases.

Calculating a Pearson Correlation Coefficient Matrix between the variable’s mpg, weight and horsepower helps to define the linear relationship the scatterplots visualize. Zybooks MAT 303: Applied Statistics II for Science Section 3.3 Table 3.3.1 defines a Strong correlation as anything above 0.80, and a Moderate Correlation between 0.40 and 0.80.

Table 3.3.1: Strength of correlation.

|  |  |
| --- | --- |
| Value of |R| | Strength of correlation |
| 0<|R|≤0.40 | Weak |
| 0.40<|R|≤0.80 | Moderate |
| 0.80<|R|≤1.00 | Strong |

The correlation matrix indicates that the Weight to MPG correlation is a strong negative correlation. While the Horsepower to MPG correlation is a moderate negative correlation.

**Table

Description automatically generated**

### Reporting Results

The general form of the Multiple Regression Model for Fuel Efficiency as the response variable (Y) with the Response Variables Weight (X1) and Horsepower (X2) is Y = β0 + β1X1 + β2X2.

The model equation of this Multiple Regression Model is Y = 37.23 – 3.88X1 – 0.03X2.

The value of R-Squared is *R-Squared* is 82.68% while the value of Adjusted R-Squared *Adjusted R-Squared* is 81.48%. R-Squared is the variation in the Dependent Variable (MPG) explained by the independent variables (Weight and Horsepower), 82.68% of variation in MPG is explained by Weight and Horsepower. Adjusted R-Squared is the variation explained by only the independent variables that significantly help in explaining the dependent variable, Adjusted R-Squared penalizes adding independent variables that do not help in predicting the dependent variable. Adjusted R-Squared is saying that, when considering the correlation of your independent variables to the dependent variables, this model accounts for 84.48% of the variation in MPG.

The beta estimates for Weight and Horsepower are -3.88 and -0.03 respectively. This indicates a negative relationship with MPG for both variables. For every unit of Weight with Horsepower fixed, you lose -3.88 MPG. For every unit of Horsepower with Weight fixed, you lose -0.03 MPG. Horsepower is less important than weight when considering Fuel Efficiency of a vehicle.

A fitted value is the value predicted by the Multiple Regression Model. The Residual is the difference between the predicted value and the actual value from data.

Chart, scatter chart

Description automatically generated

A scatterplot plotting Residual values against Fitted Values evaluates the Mean of Zero and Constance Variance assumptions of a multiple regression model. The Mean of Zero assumption assumes that the residual values in a model are zero (the response variable is linear in response to predictor variables), if they are not then the model may be nonlinear. Constant variance assumes that the residuals for each of the predictor variables should have equal or similar variance, called homoscedasticity. Heteroscedasticity is a condition where unequal and unrelated variance occurs. The scatterplot for our model does not appear to have any plot patterns obstructing linearity nor patterns in variance.

Chart, line chart, scatter chart

Description automatically generated

Above is a Q-Q plot, a Q-Q plot is a visualization of the distribution of errors. The Q stands for Quantiles, which are continuous intervals with equal probabilities across a probability distribution. A normalized set of errors will follow a linear pattern. The Q-Q plot for this model indicates normally distributed residuals.

### Evaluating Model Significance

Evaluating this model’s significance at a 5% level, we will be using the ANOVA (Analysis of Variance) F-Test. An F-test applied to a multiple regression model is used to determine if the overall model and its variables are collectively influencing the response variable in a statistically significant manner. After determining the results of the F-Test, if a linear relationship exists you can evaluate individual relationships in the model with a multiple regression individual t-test. This is useful when you want to troubleshoot or assess the value of certain predictor variables in a model.

The Overall F-Test is testing if the response variable has a linear relationship with at least one of the predictor variables. The Significance Level of our test is 5%, mathematically α = 0.05.

The Null hypothesis, H0 is that the predictor variables have no linear relationship with the response variable. Mathematically stated, the beta coefficient of predictor variables Weight and Horsepower are equal to 0.

H0 : β1 = β2 = 0

The Alternative hypothesis, Ha is that at least one predictor variable has a linear relationship with the response variable MPG. Mathematically stated, the beta coefficient of predictor variables Weight and Horsepower are not equal to 0.

Ha : βi ≠ 0 for *i* = 1, 2

The P-Value of the Overall F-Test is 9.109e-12. Since P-Value < Significance Level the null hypothesis is rejected in favor of the alternative hypothesis. At least one predictor variable has a linear relationship to the response variable MPG.

Now that we know a significant linear relationship exists with at least one of the predictor variables, we can use individual t-tests to evaluate individual relationships. We will evaluate these predictor variables at a 5% level of significance, mathematically α = 0.05.

Testing Weight with an individual t-test.

The Null Hypothesis, H0 states that the predictor variable Weight does not have a linear relationship with the response variable. Mathematically stated, the beta coefficient of weight is equal to 0.

H0 : β1 = 0

The Alternative Hypothesis, Ha states that the predictor variable Weight has a linear relationship with the response variable. Mathematically stated, the beta coefficient of Weight is not equal to 0.

Ha : β1 ≠ 0

The P-value for weight in this individual t-test is 1.12e-06. P-Value < Significance level, the Null Hypothesis is rejected. Weight has a linear relationship with the response variable MPG.

Testing Horsepower with an individual t-test.

The Null Hypothesis, H0 states that the predictor variable Horsepower does not have a linear relationship with the response variable. Mathematically stated, the beta coefficient of Horsepower is equal to 0.

H0 : β2 = 0

The Alternative Hypothesis, Ha states that the predictor variable Horsepower has a linear relationship with the response variable. Mathematically stated, the beta coefficient of Horsepower is not equal to 0.

Ha : β2 ≠ 0

The P-value for weight in this individual t-test is 0.00145. P-Value < Significance level, the Null Hypothesis is rejected. Horsepower has a linear relationship with the response variable MPG.

A confidence interval is used to measure the degree of certainty or uncertainty in a sampling method. When reading a confidence interval at 95%, the statement is “Sampling from this pool of data repeatedly, there is a 95% likelihood the data falls within this upper and lower range”. A confidence interval gauges how likely a sample contains parameters representative of the true population.

Taking a 95% confidence interval of this dataset, the 2.5% column is the upper bound and 97.5% is the lower bound.

|  | **2.5 %** | **97.5 %** |
| --- | --- | --- |
| **(Intercept)** | 33.9574 | 40.4972 |
| **wt.** | -5.1719 | -2.5837 |
| **hp** | -0.0502 | -0.0133 |

Interpreting this table,

The Weight variable will fall between a -2.5837 and -5.1719 slope coefficient. 95% of randomly drawn samples in this range will contain the true value of Weight Coefficient β1.

The Horsepower variable will fall between a -0.0133 and -0.0502 slope coefficient. 95% of randomly drawn samples in this range will contain the true value of Weight Coefficient β2.

### Making Predictions Using the Model

The predicted fuel efficiency, in MPG, for a car that has a weight of 2.95 and a horsepower of 179 is predicted to be 20.1003 by this model. If this car achieves an average of 22.7 Miles per Gallon, the residual of this observation is 22.7 – 20.1003 = 2.5997.

A prediction interval is used to predict what range a future individual observation may fall in all outcomes. A confidence interval is used to predict the range the average of the response variable will fall. In short, a prediction interval tries to account for all possible values with the model for a single observation while a confidence interval tries to predict where the average of the probability distribution will fall given specific values for the predictor variables. Confidence intervals will always be narrower than Prediction intervals since a Confidence interval is the estimation of a range of an average while the predictor interval is the range of values for a single instance.

The 95% prediction interval for the car in question is

Table

Description automatically generated

The prediction interval lower and upper ranges are 14.645 and 25.5556. If we were to observe a car with a weight of 2.95 and a horsepower of 179 once, we would observe MPG values within this lower and upper range 95% of the time based upon this model for a car with these characteristics.

The 95% confidence interval for the car in question is

Table

Description automatically generated

The confidence interval lower and upper ranges are 18.8249 to 21.3758. If we were to observe a car with a weight of 2.95 and a horsepower of 179 repeatedly, we would arrive at an average MPG value between 18.8249 and 21.3758 based on this model for a car with these characteristics; This range is predicted to be a part of the true population 95% of the time.

## Conclusion

This model predicts MPG efficiencies based upon Horsepower and Weight accounting for over 80% variance. I would not be comfortable presenting this model for use. All data used is based upon 1973-1974 models of cars. It is likely there are additional predictor variables that could be used to account for variability in MPG. It is also likely that the coefficients for mpg and horsepower have changed in the 40+ years since the 70s. Research to find additional factors and bring the model into the present is warranted. If a customer were interested in MPG ratings in the 70s, I would be comfortable recommending the model.

This multiple regression model for Fuel efficiency based upon Horsepower and Weight accounts for 82.68% of the variance in results. The multiple regression equation for this model is Y = 37.23 – 3.88X1 – 0.03X2. For every 1000 lbs. of Weight a factor of -3.88 impacts MPG. As compared to every unit of Horsepower impacting -0.03 MPG. The model was found to be statistically sound, with no violation of Mean of Zero, Constance Variance or Normalized Distribution; This indicates the data used for this model and its consequent predictions are trustworthy.

In practical terms, these analyses are useful to a wide variety of parties. A car manufacturer with the goal of increasing fuel efficiency could use this model as a starting point to set target weight and horsepower for its manufacturing. A consumer could use this model as a starting point to determine what weight and horsepower combinations to buy in a car for a fuel efficiency range. Practically speaking this model also offers room for improvement, it has ruled out any concerns that weight or horsepower do not have a linear relationship with MPG. Leading to the question, what other factors could help to predict MPG more accurately in 70s era automobiles or present automobiles?